

Guidelines for clinical trial protocols for interventions involving artificial intelligence

The SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group; Cruz Rivera, Samantha; Liu, Xiaoxuan; Chan, An-Wen; Denniston, Alastair; Calvert, Melanie

DOI:

[10.1038/s41591-020-1037-7](https://doi.org/10.1038/s41591-020-1037-7)

License:

Creative Commons: Attribution (CC BY)

Document Version

Publisher's PDF, also known as Version of record

Citation for published version (Harvard):

The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group, SPIRIT-AI and CONSORT-AI Consensus Group, Cruz Rivera, S, Liu, X, Chan, A-W, Denniston, A & Calvert, M 2020, 'Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension', *Nature Medicine*, vol. 26, no. 9, 1351–1363, pp. 1351-1363. <https://doi.org/10.1038/s41591-020-1037-7>

[Link to publication on Research at Birmingham portal](#)

General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.



OPEN

Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension

Samantha Cruz Rivera^{1,2,3}, Xiaoxuan Liu^{3,4,5,6,7}, An-Wen Chan⁸, Alastair K. Denniston^{1,3,4,5,6,9}✉, Melanie J. Calvert^{1,2,3,6,10,11,12}, The SPIRIT-AI and CONSORT-AI Working Group*, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and CONSORT-AI Consensus Group

The SPIRIT 2013 statement aims to improve the completeness of clinical trial protocol reporting by providing evidence-based recommendations for the minimum set of items to be addressed. This guidance has been instrumental in promoting transparent evaluation of new interventions. More recently, there has been a growing recognition that interventions involving artificial intelligence (AI) need to undergo rigorous, prospective evaluation to demonstrate their impact on health outcomes. The SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) extension is a new reporting guideline for clinical trial protocols evaluating interventions with an AI component. It was developed in parallel with its companion statement for trial reports: CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence). Both guidelines were developed through a staged consensus process involving literature review and expert consultation to generate 26 candidate items, which were consulted upon by an international multi-stakeholder group in a two-stage Delphi survey (103 stakeholders), agreed upon in a consensus meeting (31 stakeholders) and refined through a checklist pilot (34 participants). The SPIRIT-AI extension includes 15 new items that were considered sufficiently important for clinical trial protocols of AI interventions. These new items should be routinely reported in addition to the core SPIRIT 2013 items. SPIRIT-AI recommends that investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention will be integrated, considerations for the handling of input and output data, the human–AI interaction and analysis of error cases. SPIRIT-AI will help promote transparency and completeness for clinical trial protocols for AI interventions. Its use will assist editors and peer reviewers, as well as the general readership, to understand, interpret and critically appraise the design and risk of bias for a planned clinical trial.

A clinical trial protocol is an essential document produced by study investigators detailing a priori the rationale, proposed methods and plans for how a clinical trial will be conducted^{1,2}. This key document is used by external reviewers (funding agencies, regulatory bodies, research ethics committees, journal editors, peer reviewers, institutional review boards and, increasingly, the wider public) to understand and interpret the rationale, methodological rigor and ethical considerations of the trial. Additionally, trial protocols provide a shared reference point to support the research team in conducting a high-quality study.

Despite their importance, the quality and completeness of published trial protocols are variable^{1,2}. The SPIRIT statement was published in 2013 to provide guidance for the minimum reporting content of a clinical-trial protocol and has been widely endorsed as an international standard^{3–5}. The SPIRIT statement published in 2013 provides minimum guidance applicable for all clinical trial

interventions but recognizes that certain interventions may require extension or elaboration of these items^{1,2}. AI is an area of enormous interest, with strong drivers to accelerate new interventions through to publication, implementation and market⁶. While AI systems have been researched for some time, recent advances in deep learning and neural networks have gained considerable interest for their potential in health applications. Examples of such applications of these are wide ranging and include AI systems for screening and triage^{7,8}, diagnosis^{9–12}, prognostication^{13,14}, decision support¹⁵ and treatment recommendation¹⁶. However, in most recent cases, the majority of published evidence has consisted of *in silico*, early-phase validation. It has been recognized that most recent AI studies are inadequately reported and existing reporting guidelines do not fully cover potential sources of bias specific to AI systems¹⁷. The welcome emergence of randomized controlled trials seeking to evaluate the clinical efficacy of newer interventions based on, or including, an

¹Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK. ²Institute of Applied Health Research, University of Birmingham, Birmingham, UK. ³Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, UK. ⁴Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. ⁵University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. ⁶Health Data Research UK, London, UK. ⁷Moorfields Eye Hospital NHS Foundation Trust, London, UK. ⁸Department of Medicine, Women's College Research Institute, Women's College Hospital, University of Toronto, Ontario, Canada. ⁹National Institute of Health Research Biomedical Research Centre for Ophthalmology, Moorfields Hospital London NHS Foundation Trust and University College London, Institute of Ophthalmology, London, UK. ¹⁰National Institute of Health Research Birmingham Biomedical Research Centre, University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. ¹¹National Institute of Health Research Applied Research Collaborative West Midlands, Coventry, UK. ¹²National Institute of Health Research Surgical Reconstruction and Microbiology Centre, University of Birmingham and University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: a.denniston@bham.ac.uk

AI component (called ‘AI interventions’ here)^{15,18–23} has similarly been met with concerns about design and reporting^{17,24–26}. This has highlighted the need to provide reporting guidance that is ‘fit for purpose’ in this domain.

SPIRIT-AI (as part of the SPIRIT-AI and CONSORT-AI initiative) is an international initiative supported by SPIRIT and the EQUATOR (Enhancing the Quality and Transparency of Health Research) Network to extend or elaborate on the existing SPIRIT 2013 statement where necessary, to develop consensus-based AI-specific protocol guidance^{27,28}. It is complementary to the CONSORT-AI statement, which aims to promote high-quality reporting of AI trials. This Consensus Statement describes the methods used to identify and evaluate candidate items and gain consensus. In addition, it also provides the full SPIRIT-AI checklist, including new items and their accompanying explanations.

Methods

The SPIRIT-AI and CONSORT-AI extensions were simultaneously developed for clinical trial protocols and trial reports. An announcement for the SPIRIT-AI and CONSORT-AI initiative was published in October 2019 (ref. ²⁷), and the two guidelines were registered as reporting guidelines under development on the EQUATOR library of reporting guidelines in May 2019. Both guidelines were developed in accordance with the EQUATOR Network’s methodological framework²⁹. The SPIRIT-AI and CONSORT-AI Steering Group, consisting of 15 international experts, was formed to oversee the conduct and methodology of the study. Definitions of key terms are provided in the glossary (Box 1).

Ethical approval

This study was approved by the ethical review committee at the University of Birmingham, UK (ERN_19-1100). Participant information was provided to Delphi participants electronically before survey completion and before the consensus meeting. Delphi participants provided electronic informed consent, and written consent was obtained from consensus meeting participants.

Literature review and candidate item generation

An initial list of candidate items for the SPIRIT-AI and CONSORT-AI checklists was generated through review of the published literature and consultation with the Steering Group and known international experts. A search was performed on 13 May 2019 using the terms ‘artificial intelligence’, ‘machine learning’ and ‘deep learning’ to identify existing clinical trials for AI interventions listed within the US National Library of Medicine’s clinical trial registry (ClinicalTrials.gov). There were 316 registered trials, of which 62 were completed and 7 had published results^{22,30–35}. Two studies were reported with reference to the CONSORT statement^{22,34}, and one study provided an unpublished trial protocol³⁴. The Operations Team (X.L., S.C.R., M.J.C. and A.K.D.) identified AI-specific considerations from these studies and reframed them as candidate reporting items. The candidate items were also informed by findings from a previous systematic review that evaluated the diagnostic accuracy of deep-learning systems for medical imaging¹⁷. After consultation with the Steering Group and additional international experts ($n=19$), 29 candidate items were generated, 26 of which were relevant for both SPIRIT-AI and CONSORT-AI and 3 of which were relevant only for CONSORT-AI. The Operations Team mapped these items to the corresponding SPIRIT and CONSORT items, revising the wording and providing explanatory text as required to contextualize the items. These items were included in subsequent Delphi surveys.

Delphi consensus process

In September 2019, 169 key international experts were invited to participate in the online Delphi survey to vote upon the candidate

items and suggest additional items. Experts were identified and contacted via the Steering Group and were allowed one round of ‘snowball’ recruitment in which contacted experts could suggest additional experts. In addition, individuals who made contact following publication of the announcement were included²⁷. The Steering Group agreed that individuals with expertise in clinical trials and AI and machine learning (ML), as well as key users of the technology, should be well represented in the consultation. Stakeholders included healthcare professionals, methodologists, statisticians, computer scientists, industry representatives, journal editors, policy makers, health ‘informaticists’, experts in law and ethics, regulators, patients and funders. Participant characteristics are described in Supplementary Table 1. Two online Delphi surveys were conducted. DelphiManager software (version 4.0), developed and maintained by the COMET (Core Outcome Measures in Effectiveness Trials) initiative, was used to undertake the e-Delphi surveys. Participants were given written information about the study and were asked to provide their level of expertise within the fields of (i) AI/ML, and (ii) clinical trials. Each item was presented for consideration (26 for SPIRIT-AI and 29 for CONSORT-AI). Participants were asked to vote on each item using a 9-point scale, as follows: 1–3, not important; 4–6, important but not critical; and 7–9, important and critical. Respondents provided separate ratings for SPIRIT-AI and CONSORT-AI. There was an option to opt out of voting for each item, and each item included space for free text comments. At the end of the Delphi survey, participants had the opportunity to suggest new items. 103 responses were received for the first Delphi round, and 91 responses (88% of participants from round one) were received for the second round. The results of the Delphi surveys informed the subsequent international consensus meeting. 12 new items were proposed by the Delphi study participants and were added for discussion at the consensus meeting. Data collected during the Delphi survey were anonymized, and item-level results were presented at the consensus meeting for discussion and voting.

The two-day consensus meeting took place in January 2020 and was hosted by the University of Birmingham, UK, to seek consensus on the content of SPIRIT-AI and CONSORT-AI. 31 international stakeholders from among the Delphi survey participants were invited to discuss the items and vote on their inclusion. Participants were selected to achieve adequate representation from all the stakeholder groups. 38 items were discussed in turn, comprising the 26 items generated in the initial literature review and item-generation phase (these 26 items were relevant to both SPIRIT-AI and CONSORT-AI; 3 extra items relevant only to CONSORT-AI were also discussed) and the 12 new items proposed by participants during the Delphi surveys. Each item was presented to the consensus group, alongside its score from the Delphi exercise (median and interquartile ranges) and any comments made by Delphi participants related to that item. Consensus meeting participants were invited to comment on the importance of each item and whether the item should be included in the AI extension. In addition, participants were invited to comment on the wording of the explanatory text accompanying each item and the position of each item relative to the SPIRIT 2013 and CONSORT 2010 checklists. After open discussion of each item and the option to adjust wording, an electronic vote took place, with the option to include or exclude the item. An 80% threshold for inclusion was pre-specified and deemed reasonable by the Steering Group to demonstrate majority consensus. Each stakeholder voted anonymously using Turning Point voting pads (Turning Technologies, version 8.7.2.14).

Checklist pilot

Following the consensus meeting, attendees were given the opportunity to make final comments on the wording and agree that the updated SPIRIT-AI and CONSORT-AI items reflected discussions

Box 1 | Glossary

Artificial intelligence	The science of developing computer systems that can perform tasks normally requiring human intelligence.
AI intervention	A health intervention that relies upon an AI/ML component to serve its purpose.
CONSORT	Consolidated Standards of Reporting Trials.
CONSORT-AI extension item	An additional checklist item to address AI-specific content that is not adequately covered by CONSORT 2010.
Class-activation map	Class-activation maps are particularly relevant to image classification AI interventions. Class-activation maps are visualizations of the pixels that had the greatest influence on predicted class, by displaying the gradient of the predicted outcome from the model with respect to the input. They are also referred to as 'saliency maps' or 'heat maps'.
Health outcome	Measured variables in the trial that are used to assess the effects of an intervention.
Human-AI interaction	The process of how users (humans) interact with the AI intervention, for the AI intervention to function as intended.
Clinical outcome	Measured variables in the trial that are used to assess the effects of an intervention.
Delphi study	A research method that derives the collective opinions of a group through a staged consultation of surveys, questionnaires, or interviews, with an aim to reach consensus at the end.
Development environment	The clinical and operational settings from which the data used for training the model is generated. This includes all aspects of the physical setting (such as geographical location, physical environment), operational setting (such as integration with an electronic record system, installation on a physical device) and clinical setting (such as primary, secondary and/or tertiary care, patient disease spectrum).
Fine-tuning	Modifications or additional training performed on the AI intervention model, done with the intention of improving its performance.
Input data	The data that need to be presented to the AI intervention to allow it to serve its purpose.
Machine learning	A field of computer science concerned with the development of models/algorithms which can solve specific tasks by learning patterns from data, rather than by following explicit rules. It is seen as an approach within the field of AI.
Operational environment	The environment in which the AI intervention will be deployed, including the infrastructure required to enable the AI intervention to function.
Output data	The predicted outcome given by the AI intervention based on modeling of the input data. The output data can be presented in different forms, including a classification (including diagnosis, disease severity or stage, or recommendation such as referability), a probability, a class activation map, etc. The output data typically provide additional clinical information and/or trigger a clinical decision.
Performance error	Instances in which the AI intervention fails to perform as expected. This term can describe different types of failures, and it is up to the investigator to specify what should be considered a performance error, preferably based on prior evidence. This can range from small decreases in accuracy (compared to expected accuracy) to erroneous predictions or the inability to produce an output, in certain cases.
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials.
SPIRIT-AI	An additional checklist item to address AI-specific content that is not adequately covered by SPIRIT 2013.
SPIRIT-AI elaboration item	Additional considerations to an existing SPIRIT 2013 item when applied to AI interventions.

from the meeting. The Operations Team assigned each item as an extension or elaboration item on the basis of a decision tree and produced a penultimate draft of the SPIRIT-AI and CONSORT-AI checklists (Supplementary Fig. 1). A pilot of the penultimate checklists was conducted with 34 participants to ensure clarity of wording. Experts participating in the pilot included the following: (a) Delphi participants who did not attend the consensus meeting, and (b) external experts who had not taken part in the development process but who had reached out to the Steering Group after the Delphi study commenced. Final changes were made on wording only to improve clarity for readers, by the Operations Team (Supplementary Fig. 2).

Recommendations

SPIRIT-AI checklist items and explanation. The SPIRIT-AI extension recommends that, in conjunction with existing SPIRIT 2013 items, 15 items (12 extensions and 3 elaborations) should be addressed for trial protocols of AI interventions. These items were considered sufficiently important for clinical-trial protocols for AI interventions that they should be routinely reported in addition to the core SPIRIT 2013 checklist items. Table 1 lists the SPIRIT-AI items.

All 15 items included in the SPIRIT-AI Extension passed the threshold of 80% for inclusion at the consensus meeting. SPIRIT-AI 6a (i), SPIRIT-AI 11a (v) and SPIRIT-AI 22 each resulted from the merging of two items after discussion. SPIRIT-AI 11a (iii) did not fulfil the criteria for inclusion on the basis of its initial wording (73% vote to include); however, after extensive discussion and rewording, the consensus group unanimously supported a re-vote, at which point it passed the inclusion threshold (97% to include).

Administrative information

SPIRIT-AI 1 (i) Elaboration: Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model. *Explanation.* Indicating in the protocol title and/or abstract that the intervention involves a form of AI is encouraged, as it immediately identifies the intervention as an AI/ML intervention and also serves to facilitate indexing and searching of the trial protocol in bibliographic databases, registries and other online resources. The title should be understandable by a wide audience; therefore, a broader umbrella term such as 'artificial intelligence' or 'machine learning' is encouraged. More precise terms should be used in the abstract, rather than the title, unless they are broadly recognized as

Table 1 | SPIRIT-AI checklist

Section	Item	SPIRIT 2013 item ^a	SPIRIT-AI item		Addressed on page number ^b
Administrative information					
Title	1	Descriptive title identifying the study design, population, interventions, and, if applicable, trial acronym	SPIRIT-AI 1 (i) Elaboration	Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model.	
			SPIRIT-AI 1 (ii) Elaboration	Specify the intended use of the AI intervention.	
Trial registration	2a	Trial identifier and registry name. If not yet registered, name of intended registry			
	2b	All items from the World Health Organization Trial Registration Dataset			
Protocol version	3	Date and version identifier			
Funding	4	Sources and types of financial, material, and other support			
Roles and responsibilities	5a	Names, affiliations, and roles of protocol contributors			
	5b	Name and contact information for the trial sponsor			
	5c	Role of study sponsor and funders, if any, in study design; collection, management, analysis, and interpretation of data; writing of the report; and the decision to submit the report for publication, including whether they will have ultimate authority over any of these activities			
	5d	Composition, roles, and responsibilities of the coordinating center, steering committee, endpoint adjudication committee, data management team, and other individuals or groups overseeing the trial, if applicable (see Item 21a for data monitoring committee)			
Introduction					
Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention	SPIRIT-AI 6a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, healthcare professionals, patients, public).	
			SPIRIT-AI 6a (ii) Extension	Describe any pre-existing evidence for the AI intervention.	
	6b	Explanation for choice of comparators			
Objectives	7	Specific objectives or hypotheses			
Trial design	8	Description of trial design including type of trial (for example, parallel group, crossover, factorial, single group), allocation ratio, and framework (for example, superiority, equivalence, noninferiority, exploratory)			
Methods: participants, interventions and outcomes					
Study setting	9	Description of study settings (for example, community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained	SPIRIT-AI 9 Extension	Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting.	
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centers and individuals who will perform the interventions (for example, surgeons, psychotherapists)	SPIRIT-AI 10 (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.	
			SPIRIT-AI 10 (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.	

Continued

Table 1 | SPIRIT-AI checklist (Continued)

Section	Item	SPIRIT 2013 item ^a	SPIRIT-AI item	Addressed on page number ^b
Interventions	11a	Interventions for each group with sufficient detail to allow replication, including how and when they will be administered	SPIRIT-AI 11a (i) Extension	State which version of the AI algorithm will be used.
			SPIRIT-AI 11a (ii) Extension	Specify the procedure for acquiring and selecting the input data for the AI intervention.
			SPIRIT-AI 11a (iii) Extension	Specify the procedure for assessing and handling poor-quality or unavailable input data.
			SPIRIT-AI 11a (iv) Extension	Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users.
			SPIRIT-AI 11a (v) Extension	Specify the output of the AI intervention.
			SPIRIT-AI 11a (vi) Extension	Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice.
	11b	Criteria for discontinuing or modifying allocated interventions for a given trial participant (for example, drug dose change in response to harms, participant request, or improving/worsening disease)		
Outcomes	11c	Strategies to improve adherence to intervention protocols, and any procedures for monitoring adherence (for example, drug tablet return, laboratory tests)		
	11d	Relevant concomitant care and interventions that are permitted or prohibited during the trial		
	12	Primary, secondary, and other outcomes, including the specific measurement variable (for example, systolic blood pressure), analysis metric (for example, change from baseline, final value, time to event), method of aggregation (for example, median, proportion), and time point for each outcome. Explanation of the clinical relevance of chosen efficacy and harm outcomes is strongly recommended		
	13	Time schedule of enrollment, interventions (including any run-ins and washouts), assessments, and visits for participants. A schematic diagram is highly recommended (Fig. 1)		
Participant timeline	14	Estimated number of participants needed to achieve study objectives and how it was determined, including clinical and statistical assumptions supporting any sample size calculations		
Sample size	15	Strategies for achieving adequate participant enrollment to reach target sample size		
Recruitment		Methods: assignment of interventions (for controlled trials)		
Sequence generation	16a	Method of generating the allocation sequence (for example, computer-generated random numbers), and list of any factors for stratification. To reduce predictability of a random sequence, details of any planned restriction (for example, blocking) should be provided in a separate document that is unavailable to those who enroll participants or assign interventions		

Continued

Table 1 | SPIRIT-AI checklist (Continued)

Section	Item	SPIRIT 2013 item ^a	SPIRIT-AI item	Addressed on page number ^b
Allocation concealment mechanism	16b	Mechanism of implementing the allocation sequence (for example, central telephone; sequentially numbered, opaque, sealed envelopes), describing any steps to conceal the sequence until interventions are assigned		
Implementation	16c	Who will generate the allocation sequence, who will enroll participants, and who will assign participants to interventions		
Blinding (masking)	17a	Who will be blinded after assignment to interventions (for example, trial participants, care providers, outcome assessors, data analysts), and how		
	17b	If blinded, circumstances under which unblinding is permissible, and procedure for revealing a participant's allocated intervention during the trial		
Methods: data collection, management and analysis				
Data collection methods	18a	Plans for assessment and collection of outcome, baseline, and other trial data, including any related processes to promote data quality (for example, duplicate measurements, training of assessors) and a description of study instruments (for example, questionnaires, laboratory tests) along with their reliability and validity, if known. Reference to where data collection forms can be found, if not in the protocol		
	18b	Plans to promote participant retention and complete follow-up, including list of any outcome data to be collected for participants who discontinue or deviate from intervention protocols		
Data management	19	Plans for data entry, coding, security, and storage, including any related processes to promote data quality (for example, double data entry; range checks for data values). Reference to where details of data management procedures can be found, if not in the protocol		
Statistical methods	20a	Statistical methods for analyzing primary and secondary outcomes. Reference to where other details of the statistical analysis plan can be found, if not in the protocol		
	20b	Methods for any additional analyses (for example, subgroup and adjusted analyses)		
	20c	Definition of analysis population relating to protocol non-adherence (for example, as randomized analysis), and any statistical methods to handle missing data (for example, multiple imputation)		
Methods: monitoring				
Data monitoring	21a	Composition of data monitoring committee (DMC); summary of its role and reporting structure; statement of whether it is independent from the sponsor and competing interests; and reference to where further details about its charter can be found, if not in the protocol. Alternatively, an explanation of why a DMC is not needed		
	21b	Description of any interim analyses and stopping guidelines, including who will have access to these interim results and make the final decision to terminate the trial		

Continued

Table 1 | SPIRIT-AI checklist (Continued)

Section	Item	SPIRIT 2013 item ^a	SPIRIT-AI item	Addressed on page number ^b
Harms	22	Plans for collecting, assessing, reporting, and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct	SPIRIT-AI 22 Extension	Specify any plans to identify and analyze performance errors. If there are no plans for this, justify why not.
Auditing	23	Frequency and procedures for auditing trial conduct, if any, and whether the process will be independent from investigators and the sponsor		
Ethics and dissemination				
Research ethics approval	24	Plans for seeking research ethics committee/institutional review board (REC/IRB) approval		
Protocol amendments	25	Plans for communicating important protocol modifications (for example, changes to eligibility criteria, outcomes, analyses) to relevant parties (for example, investigators, REC/IRBs, trial participants, trial registries, journals, regulators)		
Consent or ascent	26a	Who will obtain informed consent or assent from potential trial participants or authorized surrogates, and how (see Item 32)		
	26b	Additional consent provisions for collection and use of participant data and biological specimens in ancillary studies, if applicable		
Confidentiality	27	How personal information about potential and enrolled participants will be collected, shared, and maintained in order to protect confidentiality before, during, and after the trial		
Declaration of interests	28	Financial and other competing interests for principal investigators for the overall trial and each study site		
Access to data	29	Statement of who will have access to the final trial dataset, and disclosure of contractual agreements that limit such access for investigators	SPIRIT-AI 29 Extension	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.
Ancillary and post-trial care	30	Provisions, if any, for ancillary and post-trial care, and for compensation to those who suffer harm from trial participation		
Dissemination policy	31a	Plans for investigators and sponsor to communicate trial results to participants, healthcare professionals, the public, and other relevant groups (for example, via publication, reporting in results databases, or other data sharing arrangements), including any publication restrictions		
	31b	Authorship eligibility guidelines and any intended use of professional writers		
	31c	Plans, if any, for granting public access to the full protocol, participant-level dataset, and statistical code		
Appendices				
Informed consent materials	32	Model consent form and other related documentation given to participants and authorized surrogates		
Biological specimens	33	Plans for collection, laboratory evaluation, and storage of biological specimens for genetic or molecular analysis in the current trial and for future use in ancillary studies, if applicable		

^aIt is strongly recommended that this checklist be read in conjunction with the SPIRIT 2013 Explanation & Elaboration for important clarification on the items. ^bIndicates page numbers to be completed by authors during protocol development.

being a form of AI/ML. Specific terminology relating to the model type and architecture should be detailed in the abstract.

SPIRIT-AI 1 (ii) Elaboration: State the intended use of the AI intervention. *Explanation.* The intended use of the AI intervention should be made clear in the protocol's title and/or abstract. This should describe the purpose of the AI intervention and the disease context^{19,36}. Some AI interventions may have multiple intended uses, or the intended use may evolve over time. Therefore, documenting this allows readers to understand the intended use of the algorithm at the time of the trial.

Introduction

SPIRIT-AI 6a (i) Extension: Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, healthcare professionals, patients, public). *Explanation.* In order to clarify how the AI intervention will fit into a clinical pathway, a detailed description of its role should be included in the protocol background. AI interventions may be designed to interact with different users, including healthcare professionals, patients and the public, and their roles can be wide-ranging (for example, the same AI intervention could theoretically be replacing, augmenting or adjudicating components of clinical decision-making). Clarifying the intended use of the AI intervention and its intended user helps readers understand the purpose for which the AI intervention will be evaluated in the trial.

SPIRIT-AI 6a (ii) Extension: Describe any pre-existing evidence for the AI intervention. *Explanation.* Authors should describe in the protocol any pre-existing published evidence (with supporting references) or unpublished evidence relating to validation of the AI intervention or lack thereof. Consideration should be given to whether the evidence was for a use, setting and target population similar to that of the planned trial. This may include previous development of the AI model, internal and external validations and any modifications made before the trial.

Participants, interventions and outcomes

SPIRIT-AI 9 Extension: Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting. *Explanation.* There are limitations to the generalizability of AI algorithms, one of which is when they are used outside of their development environment^{37,38}. AI systems are dependent on their operational environment, and the protocol should provide details of the hardware and software requirements to allow technical integration of the AI intervention at each study site. For example, it should be stated if the AI intervention requires vendor-specific devices, if there is a need for specialized computing hardware at each site, or if the sites must support cloud integration, particularly if this is vendor specific. If any changes to the algorithm are required at each study site as part of the implementation procedure (such as fine-tuning the algorithm on local data), then this process should also be clearly described.

SPIRIT-AI 10 (i) Elaboration: State the inclusion and exclusion criteria at the level of participants. *Explanation.* The inclusion and exclusion criteria should be defined at the participant level as per usual practice in protocols of non-AI interventional trials. This is distinct from the inclusion and exclusion criteria made at the input data level, which are addressed in item 10 (ii).

SPIRIT-AI 10 (ii) Extension: State the inclusion and exclusion criteria at the level of the input data. *Explanation.* 'Input data' refers to the data required by the AI intervention to serve its purpose (for example, for a breast cancer diagnostic system, the input

data could be the unprocessed or vendor-specific post-processing mammography scan upon which a diagnosis is being made; for an early-warning system, the input data could be physiological measurements or laboratory results from the electronic health record). The trial protocol should pre-specify if there are minimum requirements for the input data (such as image resolution, quality metrics or data format) that would determine pre-randomization eligibility. It should specify when, how and by whom this will be assessed. For example, if a participant met the eligibility criteria for lying flat for a CT scan as per item 10 (i), but the scan quality was compromised (for any given reason) to such a level that it is no longer fit for use by the AI system, this should be considered as an exclusion criterion at the input-data level. Note that where input data are acquired after randomization (addressed by SPIRIT-20c), any exclusion is considered to be from the analysis, not from enrollment (Fig. 1).

SPIRIT-AI 11a (i) Extension: State which version of the AI algorithm will be used. *Explanation.* Similar to other forms of software as a medical device, AI systems are likely to undergo multiple iterations and updates in their lifespan. The protocol should state which version of the AI system will be used in the clinical trial and whether this is the same version that was used in previous studies that have been used to justify the study rationale. If applicable, the protocol should describe what has changed between the relevant versions and the rationale for the changes. Where available, the protocol should include a regulatory marking reference, such as a unique device identifier, that requires a new identifier for updated versions of the device³⁹.

SPIRIT-AI 11a (ii) Extension: Specify the procedure for acquiring and selecting the input data for the AI intervention. *Explanation.* The measured performance of any AI system may be critically dependent on the nature and quality of the input data⁴⁰. The procedure for how input data will be handled, including data acquisition, selection and pre-processing before analysis by the AI system, should be provided. Completeness and transparency of this process is integral to feasibility assessment and to future replication of the intervention beyond the clinical trial. It will also help to identify whether input-data-handling procedures will be standardized across trial sites.

SPIRIT-AI 11a (iii) Extension: Specify the procedure for assessing and handling poor-quality or unavailable input data. *Explanation.* As with SPIRIT-AI 10 (ii), 'input data' refers to the data required by the AI intervention to serve its purpose. As noted in item 10 (ii), the performance of AI systems may be compromised as a result of poor quality or missing input data⁴¹ (for example, excessive movement artifact on an electrocardiogram). The study protocol should specify if and how poor quality or unavailable input data will be identified and handled. The protocol should also specify a minimum standard required for the input data and the procedure for when the minimum standard is not met (including the impact on, or any changes to, the participant care pathway).

Poor quality or unavailable data can also affect non-AI interventions. For example, sub-optimal quality of a scan could affect a radiologist's ability to interpret it and make a diagnosis. It is therefore important that this information is reported equally for the control intervention, where relevant. If this minimum quality standard is different from the inclusion criteria for input data used to assess eligibility pre-randomization, this should be stated.

SPIRIT-AI 11a (iv) Extension: Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users. *Explanation.* A description of the human-AI interface and the requirements for successful interaction when input data are handled should be provided. Examples include

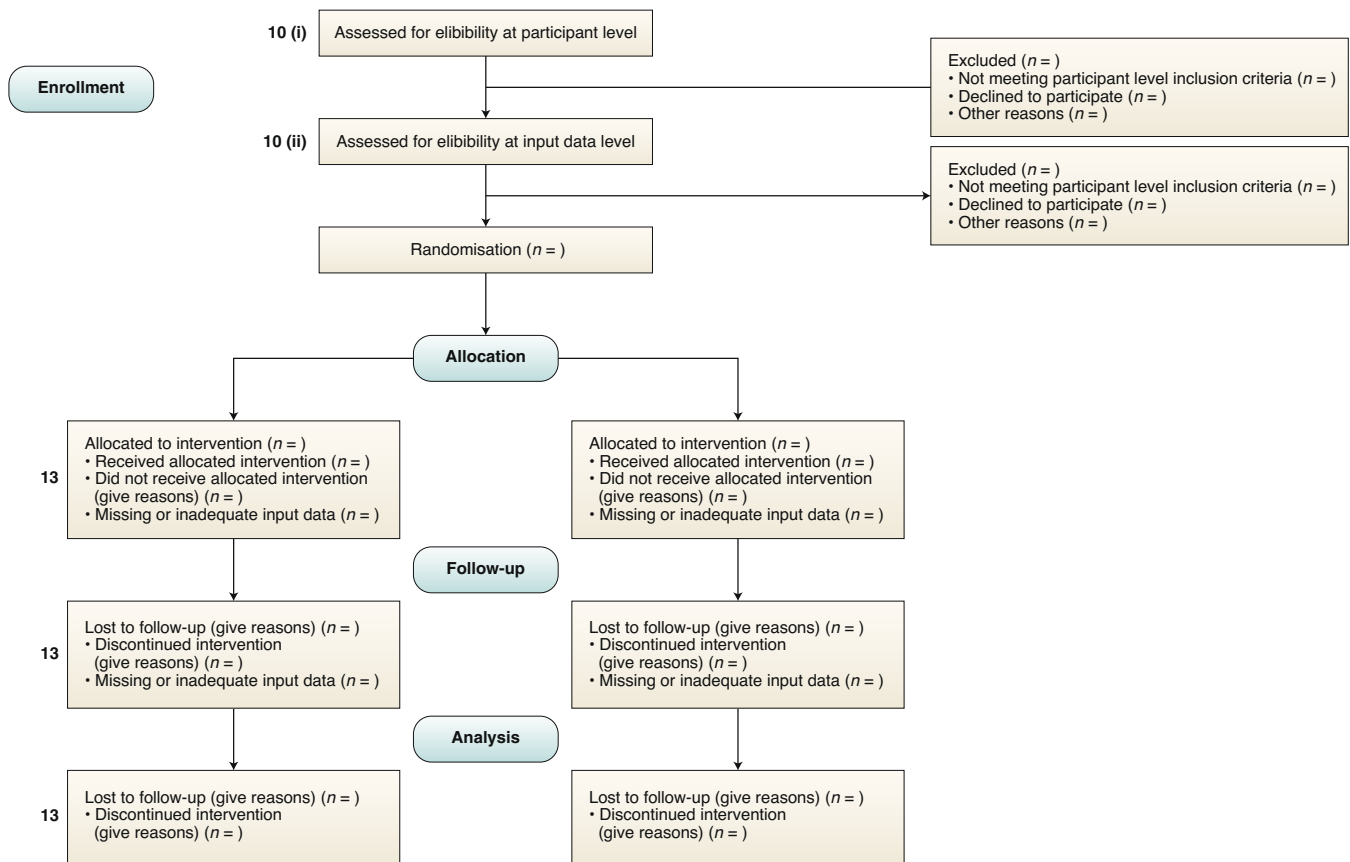


Fig. 1 | CONSORT 2010 flow diagram — adapted for AI clinical trials. SPIRIT-AI 10 (i): State the inclusion and exclusion criteria at the level of participants. SPIRIT-AI 10 (ii): State the inclusion and exclusion criteria at the level of the input data. SPIRIT 13 (core CONSORT item): Time schedule of enrollment, interventions (including any run-ins and washouts), assessments, and visits for participants. A schematic diagram is highly recommended.

clinician-led selection of regions of interest from a histology slide that is then interpreted by an AI diagnostic system⁴², or an endoscopist's selection of a colonoscopy video clips as input data for an algorithm designed to detect polyps²¹. A description of any planned user training and instructions for how users will handle the input data provides transparency and replicability of trial procedures. Poor clarity on the human–AI interface may lead to a lack of a standard approach and may carry ethical implications, particularly in the event of harm^{43,44}. For example, it may become unclear whether an error case occurred due to human deviation from the instructed procedure, or if it was an error made by the AI system.

SPIRIT-AI 11a (v) Extension: Specify the output of the AI intervention. *Explanation.* The output of the AI intervention should be clearly defined in the protocol. For example, an AI system may output a diagnostic classification or probability, a recommended action, an alarm alerting to an event, an instigated action in a closed-loop system (such as titration of drug infusions) or another output. The nature of the AI intervention's output has direct implications on its usability and how it may lead to downstream actions and outcomes.

SPIRIT-AI 11a (vi) Extension: Explain the procedure for how the AI intervention's outputs will contribute to decision-making or other elements of clinical practice. *Explanation.* Since health outcomes may also critically depend on how humans interact with the AI intervention, the trial protocol should explain how the outputs of the AI system are used to contribute to decision-making or other elements of clinical practice. This should include adequate

description of downstream interventions that can impact outcomes. As with SPIRIT-AI 11a (iv), any effects of human–AI interaction on the outputs should be described in detail, including the level of expertise required to understand the outputs and any training and/or instructions provided for this purpose. For example, a skin cancer detection system that produces a percentage likelihood as output should be accompanied by an explanation of how this output should be interpreted and acted upon by the user, specifying both the intended pathways (for example, skin lesion excision if the diagnosis is positive) and the thresholds for entry to these pathways (for example, skin lesion excision if the diagnosis is positive and the probability is >80%). The information produced by comparator interventions should be similarly described, alongside an explanation of how such information was used to arrive at clinical decisions for patient management, where relevant.

Monitoring

SPIRIT-AI 22 Extension: Specify any plans to identify and analyze performance errors. If there are no plans for this, explain why not. *Explanation.* Reporting performance errors and failure case analysis is especially important for AI interventions. AI systems can make errors that may be hard to foresee but that, if allowed to be deployed at scale, could have catastrophic consequences⁴⁵. Therefore, identifying cases of error and defining risk-mitigation strategies is important for informing when the intervention can be safely implemented, and for which populations. The protocol should specify whether there are any plans to analyze performance errors. If there are no plans for this, a justification should be included in the protocol.

Ethics and dissemination

SPIRIT-AI 29 Extension: State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use. *Explanation.* The protocol should make clear whether and how the AI intervention and/or its code can be accessed or re-used. This should include details about the license and any restrictions to access.

Discussion

The SPIRIT-AI extension provides international consensus-based guidance on AI-specific information that should be reported in clinical trial protocols, alongside SPIRIT 2013 and other relevant SPIRIT extensions^{4,46}. It comprises of 15 items: 3 elaborations to the existing SPIRIT 2013 guidance in the context of AI trials, and 12 new extensions. The guidance does not aim to be prescriptive about the methodological approach to AI trials; instead, it aims to promote transparency in reporting the design and methods of a clinical trial to facilitate understanding, interpretation and peer review.

A number of extension items relate to the intervention (items 11 (i)–11 (vi)), its setting (item 9) and intended role (item 6a (i)). Specific recommendations were made pertinent to AI systems related to algorithm version, input and output data, integration into trial settings, expertise of the users and protocol for acting upon the AI system's recommendations. It was agreed that these details are critical for independent evaluation of the study protocol. Journal editors reported that despite the importance of these items, they are currently often missing from trial protocols and reports at the time of submission for publication, which provides further weight to their inclusion as specifically listed extension items.

A recurrent focus of the Delphi comments and consensus group discussion was the safety of AI systems. This is in recognition that these systems, unlike other health interventions, can unpredictably yield errors that are not easily detectable or explainable by human judgement. For example, changes to medical imaging that are invisible, or appear random, to the human eye may change the likelihood of the resultant diagnostic output entirely^{47,48}. The concern is that given the theoretical ease with which AI systems could be deployed at scale, any unintended harmful consequences could be catastrophic. Two extension items were added to address this. SPIRIT-AI item 6a (ii) requires specification of the prior level of evidence for validation of the AI intervention. SPIRIT-AI item 22 requires specification of any plans to analyze performance errors, to emphasize the importance of anticipating systematic errors made by the algorithm and their consequences.

One topic that was raised in the Delphi survey responses and consensus meeting that is not included in the final guidelines is 'continuously evolving' AI systems (also known as 'continuously adapting' or 'continuously learning' AI systems). These are AI systems with the ability to continuously train on new data, which may cause changes in performance over time. The group noted that, while of interest, this field is relatively early in its development without tangible examples in healthcare applications, and that it would not be appropriate for it to be addressed by SPIRIT-AI at this stage⁴⁹. This topic will be monitored and revisited in future iterations of SPIRIT-AI. It is worth noting that incremental software changes, whether continuous or iterative, intentional or unintentional, could have serious consequences on safety performance after deployment. It is therefore of vital importance that such changes are documented and identified by software version and that a robust post-deployment surveillance plan is in place.

This study is set in the current context of AI in health; therefore, several limitations should be noted. First, at the time of SPIRIT-AI development, there were only seven published trials and no published trial protocols in the field of AI for healthcare. Thus, the

discussion and decisions made during the development of SPIRIT-AI are not always supported by existing real-world examples. This arises from our stated aim of addressing the issues of poor protocol development in this field as early as possible, recognizing the strong drivers in the field and the specific challenges of study design and reporting for AI. As the science and study of AI evolves, we welcome collaboration with investigators to co-evolve these reporting standards to ensure their continued relevance. Second, the literature search of AI randomized controlled trials used terminology such as 'artificial intelligence', 'machine learning' and 'deep learning', but not terms such as 'clinical decision support systems' and 'expert systems', which were more commonly used in the 1990s for technologies underpinned by AI systems and share risks similar to those of recent examples⁵⁰. It is likely that such systems, if published today, would be indexed under 'artificial intelligence' or 'machine learning'; however, clinical decision support systems were not actively discussed during this consensus process. Third, the initial candidate items list was generated by a relatively small group of experts consisting of Steering Group members and additional international experts. However, additional items from the wider Delphi group were taken forward for consideration by the consensus group, and no new items were suggested during the consensus meeting or post-meeting evaluation.

As with the SPIRIT statement, the SPIRIT-AI extension is intended as a minimum reporting guidance, and there are additional AI-specific considerations for trial protocols that may warrant consideration (Supplementary Table 2). This extension is aimed particularly at investigators planning or conducting clinical trials; however, it may also serve as useful guidance for developers of AI interventions in earlier validation stages of an AI system. Investigators seeking to report studies developing and validating the diagnostic and predictive properties of AI models should refer to TRIPOD-ML (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis–Machine Learning)²⁴ and STARD-AI (Standards For Reporting Diagnostic Accuracy Studies–Artificial Intelligence)⁵¹, both of which are currently under development. Other potentially relevant guidelines, which are agnostic to study design, are registered with the EQUATOR network⁵². The SPIRIT-AI extension is expected to encourage careful early planning of AI interventions for clinical trials and this, in conjunction with CONSORT-AI, should help to improve the quality of trials for AI interventions.

There is widespread recognition that AI is a rapidly evolving field, and there will be the need to update SPIRIT-AI as the technology, and newer applications for it, develop. Currently, most applications of AI/ML involve disease detection, diagnosis and triage, and this is likely to have influenced the nature and prioritization of items within SPIRIT-AI. As wider applications that utilize 'AI as therapy' emerge, it will be important to re-evaluate SPIRIT-AI in the light of such studies. Additionally, advances in computational techniques and the ability to integrate them into clinical workflows will bring new opportunities for innovation that benefits patients. However, they may be accompanied by new challenges of study design and reporting to ensure transparency, minimize potential biases and ensure that the findings of such a study are trustworthy and the extent to which they may be generalizable. The SPIRIT-AI and CONSORT-AI Steering Group will continue to monitor the need for updates.

Data availability

Data requests should be made to the corresponding author and release will be subject to consideration by the SPIRIT-AI and CONSORT-AI Steering Group.

Received: 24 April 2020; Accepted: 23 July 2020;
Published online: 9 September 2020

References

- Chan, A.-W. et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann. Intern. Med.* **158**, 200–207 (2013).
- Chan, A.-W. et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *Br. Med. J.* **346**, e7586 (2013).
- Sarkis-Onofre, R. et al. Use of guidelines to improve the quality and transparency of reporting oral health research. *J. Dent.* **43**, 397–404 (2015).
- Calvert, M. et al. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: The SPIRIT-PRO Extension. *J. Am. Med. Assoc.* **319**, 483–494 (2018).
- Dai, L. et al. Standard protocol items for clinical trials with traditional Chinese medicine 2018: recommendations, explanation and elaboration (SPIRIT-TCM Extension 2018). *Chin. J. Integr. Med.* **25**, 71–79 (2019).
- He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
- McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
- Abramoff, M. D. et al. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest. Ophthalmol. Vis. Sci.* **57**, 5200–5206 (2016).
- De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
- Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
- Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
- Fleuren, L. M. et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.* **46**, 383–400 (2020).
- Yim, J. et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat. Med.* **26**, 892–899 (2020).
- Kim, H., Goo, J. M., Lee, K. H., Kim, Y. T. & Park, C. M. Preoperative CT-based deep learning model for predicting disease-free survival in patients with lung adenocarcinomas. *Radiology* **296**, 216–224 (2020).
- Wang, P. et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813–1819 (2019).
- Tyler, N. S. et al. An artificial intelligence decision support system for the management of type 1 diabetes. *Nat. Metab.* **2**, 612–619 (2020).
- Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
- Wu, L. et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* **68**, 2161–2169 (2019).
- Wijnberge, M. et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: The HYPE randomized clinical trial. *J. Am. Med. Assoc.* **323**, 1052–1060 (2020).
- Gong, D. et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol. Hepatol.* **5**, 352–361 (2020).
- Wang, P. et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADE-DB trial): a double-blind randomised study. *Lancet Gastroenterol. Hepatol.* **5**, 343–351 (2020).
- Lin, H. et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* **9**, 52–59 (2019).
- Su, J.-R. et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). *Gastrointest. Endosc.* **91**, 415–424.e4 (2020).
- Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
- Gregory, J., Welliver, S. & Chong, J. Top 10 reviewer critiques of radiology artificial intelligence (AI) articles: qualitative thematic analysis of reviewer critiques of machine learning/deep learning manuscripts submitted to JMIR. *J. Magn. Reson. Imaging* **52**, 248–254 (2020).
- Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Br. Med. J.* **368**, m689 (2020).
- CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med.* **25**, 1467–1468 (2019).
- Liu, X., Faes, L., Calvert, M. J. & Denniston, A. K. Extension of the CONSORT and SPIRIT statements. *Lancet* **394**, 1225 (2019).
- Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLoS Med.* **7**, e1000217 (2010).
- Caballero-Ruiz, E. et al. A web-based clinical decision support system for gestational diabetes: Automatic diet prescription and detection of insulin needs. *Int. J. Med. Inform.* **102**, 35–49 (2017).
- Kim, T. W. B., Gay, N., Khemka, A. & Garino, J. Internet-based exercise therapy using algorithms for conservative treatment of anterior knee pain: a pragmatic randomized controlled trial. *JMIR Rehabil. Assist. Technol.* **3**, e12 (2016).
- Labovitz, D. L., Shafner, L., Reyes Gil, M., Virmani, D. & Hanina, A. Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy. *Stroke* **48**, 1416–1419 (2017).
- Nicolae, A. et al. Evaluation of a machine-learning algorithm for treatment planning in prostate low-dose-rate brachytherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **97**, 822–829 (2017).
- Voss, C. et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatr.* **173**, 446–454 (2019).
- Mendes-Soares, H. et al. Assessment of a personalized approach to predicting postprandial glycemic responses to food among individuals without diabetes. *JAMA Netw. Open* **2**, e188102 (2019).
- Choi, K. J. et al. Development and validation of a deep learning system for staging liver fibrosis by using contrast agent-enhanced CT images in the liver. *Radiology* **289**, 688–697 (2018).
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
- Pooch, E. H. P., Ballester, P. L. & Barros, R. C. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv* <http://arxiv.org/abs/1909.01940> (2019).
- International Medical Device Regulators Forum. Unique device identification system (UDI System) application guide. <http://www.imdrf.org/documents/documents.asp> (2019).
- Sabottke, C. F. & Spieler, B. M. The effect of image resolution on deep learning in radiography. *Radiology: Artif. Intell.* **2**, e190015 (2020).
- Heaven, D. Why deep-learning AIs are so easy to fool. *Nature* **574**, 163–166 (2019).
- Kiani, A. et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit. Med.* **3**, 23 (2020).
- Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
- Habli, I., Lawton, T. & Porter, Z. Artificial intelligence in health care: accountability and safety. *Bull. World Health Organ.* https://www.who.int/bulletin/online_first/BLT.19.237487.pdf (2020).
- Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *arXiv* <http://arxiv.org/abs/1909.12475> (2019).
- SPIRIT. Publications & Downloads. <https://www.spirit-statement.org/publications-downloads/> (accessed 24 March 2020).
- Zech, J. R. et al. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv* <http://arxiv.org/abs/1807.00431> (2018).
- Finlayson, S. G. et al. Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
- Lee, C. S. & Lee, A. Y. Clinical applications of continual learning machine learning. *Lancet Digit. Health* **2**, e279–e281 (2020).
- Sutton, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit. Med.* **3**, 17 (2020).
- Sounderajah, V. et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat. Med.* **26**, 807–808 (2020).
- Talmon, J. et al. STARE-HI—Statement on reporting of evaluation studies in Health Informatics. *Int. J. Med. Inform.* **78**, 1–9 (2009).

Acknowledgements

We thank the participants who were involved in the Delphi study and Pilot study (Supplementary Note); E. Marston (University of Birmingham, UK) for strategic support; and C. Radovanovic (University Hospitals Birmingham NHS Foundation Trust, UK) and A. Walker (University of Birmingham, UK) for administrative support. The views expressed in this publication are those of the authors, Delphi participants and stakeholder participants and may not represent the views of the broader stakeholder group or host institution. This work was funded by a Wellcome Trust Institutional Strategic Support Fund: Digital Health Pilot Grant Research England (part of UK Research and Innovation), Health Data Research UK and the Alan Turing Institute. The study was sponsored by the University of Birmingham, UK. The study funders and sponsors had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; or decision to submit the manuscript for publication. M.J.C. is a National Institute for Health Research (NIHR) Senior Investigator and receives funding from the National Institute for Health Research (NIHR) Birmingham Biomedical Research Centre; the NIHR Surgical Reconstruction and Microbiology Research Centre and NIHR ARC West

Midlands at the University of Birmingham and University Hospitals Birmingham NHS Foundation Trust; Health Data Research UK; Innovate UK (part of UK Research and Innovation); the Health Foundation; Macmillan Cancer Support; and UCB Pharma. A.D. and J.D. are also NIHR Senior Investigators. The views expressed in this article are those of the author(s) and not necessarily those of the NIHR, or the Department of Health and Social Care. S.J.V. receives funding from the Engineering and Physical Sciences Research Council, UK Research and Innovation (UKRI), Accenture, Warwick Impact Fund, Health Data Research UK and European Regional Development Fund. S.R. is an employee of the Medical Research Council (UKRI). D.M. is supported by a University of Ottawa Research Chair. A.B. is supported by a National Institutes of Health (NIH) award (7K01HL141771-02). M.K.E. is supported by the U.S. Food and Drug Administration (FDA), and D.P. is supported in part by the Office of the Director at the National Library of Medicine (NLM), US National Institutes of Health (NIH). This article may not be consistent with NIH and/or FDA's views or policies. It reflects only the views and opinions of the authors.

Author contributions

Concept and design, and acquisition, analysis and interpretation of data, all authors; drafting of the manuscript, X.L., S.C.R., A.W.C., M.J.C. and A.K.D.; obtaining of funding, A.K.D., M.J.C., C.Y. and C.H. The SPIRIT-AI and CONSORT-AI Working Group consists of two groups that have been key in the development of the guidelines: the SPIRIT-AI and CONSORT-AI Steering Group, which was responsible for overseeing the consensus process and guidelines development methodology (Alastair K. Denniston, An-Wen Chan, Ara Darzi, Christopher Holmes, Christopher Yau, David Moher, Hutan Ashrafian, Jonathan J. Deeks, Lavinia Ferrante di Ruffano, Livia Faes, Melanie J. Calvert, Pearse A. Keane, Samantha Cruz Rivera, Sebastian J. Vollmer and Xiaoxuan Liu); and the SPIRIT-AI and CONSORT-AI Consensus Group, which was responsible for reaching consensus on the content and wording of the items within the checklists (Aaron Y. Lee, Adrian Jonas, Andre Esteve, Andrew L. Beam, An-Wen Chan, Maria Beatrice Panico, Cecilia S. Lee, Charlotte Haug, Christopher J. Kelly, Christopher Yau, Cynthia Mulrow, Cyrus Espinoza, David Moher, Dina Paltoo, Elaine Manna, Gary Price, Gary S. Collins, Hugh Harvey, James Matcham, Joao Monteiro, John Fletcher, M. Khair ElZarrad, Lavinia Ferrante di Ruffano, Luke Oakden-Rayner, Melanie J. Calvert, Melissa McCradden, Pearse A. Keane, Richard Savage, Robert Golub, Rupa Sarkar and Samuel Rowley).

Competing interests

M.J.C. has received personal fees from Astellas, Takeda, Merck, Daiichi Sankyo, Glaukos, GlaxoSmithKline and the Patient-Centered Outcomes Research Institute (PCORI) outside the submitted work. P.A.K. is a consultant for DeepMind Technologies, Roche, Novartis and Apellis, and has received speaker fees or travel support from Bayer, Allergan, Topcon and Heidelberg Engineering. C.J.K. is an employee of Google and owns Alphabet stock. A.E. is an employee of Salesforce CRM. R.S. is an employee of Pinpoint Science. J. Matcham was an employee of AstraZeneca at the time of this study. J. Monteiro is Chief Editor of the journal *Nature Medicine*; he has recused himself from any aspect of decision-making on this manuscript and played no part in the assignment of this manuscript to in-house editors or peer reviewers, and was also separated and blinded from the editorial process from submission inception to decision.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41591-020-1037-7>.

Correspondence and requests for materials should be addressed to A.K.D.

Peer review information Javier Carmona was the primary editor on this article, and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

SPIRIT-AI and CONSORT-AI Steering Group

Alastair K. Denniston^{1,3,4,5,6,9}, An-Wen Chan⁸, Ara Darzi^{13,14}, Christopher Holmes^{15,16}, Christopher Yau^{15,17}, David Moher^{18,19}, Hutan Ashrafian^{13,14}, Jonathan J. Deeks^{2,10}, Lavinia Ferrante di Ruffano², Livia Faes²⁰, Melanie J. Calvert^{1,2,3,6,10,11,12}, Pearse A. Keane¹, Samantha Cruz Rivera^{1,2,3}, Sebastian J. Vollmer^{15,21} and Xiaoxuan Liu^{3,4,5,6,7}

¹³Patient Safety Translational Research Centre, Imperial College London, London, UK. ¹⁴Institute of Global Health Innovation, Imperial College London, London, UK. ¹⁵Alan Turing Institute, London, UK. ¹⁶Department of Statistics and Nuffield Department of Medicine, University of Oxford, Oxford, UK. ¹⁷University of Manchester, Manchester, UK. ¹⁸Centre for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada. ¹⁹School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada. ²⁰Department of Ophthalmology, Cantonal Hospital Lucerne, Lucerne, Switzerland. ²¹University of Warwick, Coventry, UK.

SPIRIT-AI and CONSORT-AI Consensus Group

Aaron Y. Lee²², Adrian Jonas²³, Andre Esteve²⁴, Andrew L. Beam²⁵, An-Wen Chan⁸, Maria Beatrice Panico²⁶, Cecilia S. Lee²², Charlotte Haug²⁷, Christophe J. Kelly²⁸, Christopher Yau^{15,17}, Cynthia Mulrow²⁹, Cyrus Espinoza³⁰, John Fletcher³¹, David Moher^{18,19}, Dina Paltoo³², Elaine Manna³³, Gary Price³⁴, Gary S. Collins³⁵, Hugh Harvey³⁶, James Matcham³⁷, Joao Monteiro³⁸, M. Khair ElZarrad³⁹, Lavinia Ferrante di Ruffano², Luke Oakden-Rayner⁴⁰, Melanie J. Calvert^{1,2,3,6,10,11,12}, Melissa McCradden⁴¹, Pearse A. Keane¹, Richard Savage⁴², Robert Golub⁴³, Rupa Sarkar⁴⁴ and Samuel Rowley⁴⁵

²²Department of Ophthalmology, University of Washington, Seattle, WA, USA. ²³The National Institute for Health and Care Excellence, London, UK.

²⁴Salesforce Research, San Francisco, CA, USA. ²⁵Harvard T.H. Chan School of Public Health, Boston, MA, USA. ²⁶Medicines and Healthcare products

Regulatory Agency, London, UK. ²⁷New England Journal of Medicine, Waltham, MA, USA. ²⁸Google Health, London, UK. ²⁹Annals of Internal Medicine, Philadelphia, PA, USA. ³⁰Patient Partner, Birmingham, UK. ³¹British Medical Journal, London, UK. ³²National Institutes of Health, Bethesda, MD, USA. ³³Patient Partner, London, UK. ³⁴Patient Partner, Centre for Patient Reported Outcome Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK. ³⁵Centre for Statistics in Medicine, University of Oxford, Oxford, UK. ³⁶Hardian Health, London, UK. ³⁷AstraZeneca, Cambridge, UK. ³⁸Nature Research, New York, NY, USA. ³⁹Food and Drug Administration, Silver Spring, MD, USA. ⁴⁰Australian Institute for Machine Learning, North Terrace, Adelaide, Australia. ⁴¹The Hospital for Sick Children, Toronto, Canada. ⁴²PinPoint Data Science, Leeds, UK. ⁴³Journal of the American Medical Association, Chicago, IL, USA. ⁴⁴The Lancet Group, London, UK. ⁴⁵Medical Research Council, London, UK.